Maximum likelihood kernel density estimation: On the potential of convolution sieves

Accepted Manuscript

Maximum likelihood kernel density estimation: On the potential of convolution sieves

M.C. Jones, D.A. Henderson

PII: S0167-9473(09)00126-1
DOI: 10.1016/j.csda.2009.03.019
Reference: COMSTA 4348

To appear in: *Computational Statistics and Data Analysis*

Received date: 6 December 2007
Revised date: 13 January 2009
Accepted date: 24 March 2009

Please cite this article as: Jones, M.C., Henderson, D.A., Maximum likelihood kernel density estimation: On the potential of convolution sieves. *Computational Statistics and Data Analysis* (2009), doi:10.1016/j.csda.2009.03.019

Maximum likelihood kernel density estimation: on the potential of convolution sieves

M.C. Jones[a,*] and D.A. Henderson[b]

[a] *Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK*

[b] *School of Mathematics and Statistics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK*

Abstract

Methods for improving the basic kernel density estimator include variable locations, variable bandwidths and variable weights. Typically these methods are implemented separately and via pilot estimation of variation functions derived from asymptotic considerations. The starting point here is a simple maximum likelihood procedure which allows (in its greatest generality) variation of all these quantities at once, bypassing asymptotics and explicit pilot estimation. One special case of this approach is the density estimator associated with nonparametric maximum likelihood estimation (NPMLE) in a normal location mixture model. Another, closely associated with the NPMLE, is a kernel convolution sieve estimator proposed in 1982 but little used in practice to date. Simple algorithms are utilised, a simulation study is reported on, a method for bandwidth selection is investigated and an illustrative example is given. The simulations and other considerations suggest that the kernel convolution sieve provides an especially promising framework for further practical utilisation and development. And the method has a further advantage: it automatically reduces, where appropriate, to a few-component mixture model which indicates and initialises parametric mixture modelling of the data.

Keywords: Mixture modelling; Normal mixtures; NPMLE; Variable location.

*Corresponding author. Tel: +44 1908 652209; fax: +44 1908 655515.
E-mail address: m.c.jones@open.ac.uk (M.C. Jones)

1. Introduction

Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from a univariate distribution with unknown density f. Let K be a symmetric probability density function to be used as a kernel and $h > 0$ its scaling parameter, or bandwidth; write $K_h(\cdot) = h^{-1}K(h^{-1}\cdot)$. Then the standard kernel estimator of the density f at a point x is given by

$$\widehat{f}(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)$$

(Silverman, 1986, Wand and Jones, 1995, Simonoff, 1996). Let $\mathbf{m} = (m_1, ..., m_n)$, $\mathbf{w} = (w_1, ..., w_n)$ and $\mathbf{b} = (b_1, ..., b_n)$ with $w_i \geq 0$ and $b_i > 0$, $i = 1, ..., n$. One interpretation of the kernel density estimator is as a special case of an over-parametrised mixture model of the form

$$\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}(x) = \left\{\sum_{i=1}^{n} w_i\right\}^{-1} \sum_{i=1}^{n} w_i K_{b_i}(x - m_i) \tag{1}$$

where each datapoint X_i is associated with its own mixture component density K with its own location, m_i, scale parameter or bandwidth, b_i, and weight, w_i. Clearly, $\widehat{f}(x) = \widehat{f}_{\mathbf{X},\mathbf{1},h\mathbf{1}}(x)$ where $\mathbf{1} = (1, ..., 1)$. Other special cases of the general formulation include variable location kernel density estimators of the form $\widehat{f}_{\mathbf{m}}(x) \equiv \widehat{f}_{\mathbf{m},\mathbf{1},h\mathbf{1}}(x)$ (Samiuddin and el-Sayyad, 1990, called 'data sharpening' by Choi and Hall, 1999), variable weight kernel density estimators $\widehat{f}_{\mathbf{w}}(x) \equiv \widehat{f}_{\mathbf{X},\mathbf{w},h\mathbf{1}}(x)$ (Hall and Turlach, 1999) and variable bandwidth kernel density estimators $\widehat{f}_{\mathbf{b}}(x) \equiv \widehat{f}_{\mathbf{X},\mathbf{1},\mathbf{b}}(x)$ (Abramson, 1982), the last often being referred to as variable kernel density estimators. These approaches can all achieve 'higher order bias' and as such most of them are reviewed and compared with other higher order bias kernel density estimators in Jones and Signorini (1997).

In an earlier technical report (Jones and Henderson, 2005), we simply fitted the general model $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}(x)$ and each of its submodels such as $\widehat{f}_{\mathbf{m}}(x)$ or perhaps $\widehat{f}_{\mathbf{m},\mathbf{b}}(x) \equiv \widehat{f}_{\mathbf{m},\mathbf{1},\mathbf{b}}(x)$ to data by maximum likelihood, with the important proviso that $g(\mathbf{b}) \equiv (\prod_{i=1}^{n} b_i)^{1/n} = h$ for a given overall bandwidth h. In one case, $\widehat{f}_{\mathbf{m},\mathbf{w}}$, we were actually utilising the nonparametric maximum likelihood estimator (NPMLE; Laird, 1978, Lindsay, 1983, 1995) of a normal location mixture model as density estimator. We note that, while there has been much interest in normal mixture NPMLEs, most of that interest

has centred on the estimation of the mixture distribution itself and little on the resulting mixture density thought of as a density estimate *per se*. We found empirically that the best of all these estimators can achieve every bit as good performance as the best of the higher order bias kernel density estimators. Moreover, they achieve this level of performance without resort to the asymptotic arguments that underlie the practical implementation of most higher order bias kernel density estimators (Jones and Signorini, 1997).

In a fairly substantial simulation study to which the reader is referred for comparative details, Jones and Henderson (2005) argued that the best of these estimators is $\widehat{f}_{\mathbf{m}}$ and so it is this estimator that we concentrate on in this short article. This particular estimator turns out to be precisely the (normal) kernel convolution sieve estimator proposed in Section 6 of Geman and Hwang (1982; $\widehat{S}_m$ in their notation). After describing it, we provide empirical evidence of its considerable potential in results extracted from our earlier simulations. We also briefly describe the least squares cross-validation method for selection of h but have to report disappointing empirical results for it. We have been unable to contribute to the difficult problem of the theoretical analysis of the performance of $\widehat{f}_{\mathbf{m}}$ as an estimator of f (Ghosal and van der Vaart, 2001, and references therein). We provide application to a real data example which illustrates how (as also mentioned in Section 2) the well known sparseness properties of NPMLE solutions transfer to $\widehat{f}_{\mathbf{m}}$, whose form can then be used to suggest a much simpler parametric finite mixture model for the data.

In many ways, this article is an encomium to the potential of the convolution sieve approach to kernel-type density estimation which has heretofore remained to some extent an object of theoretical scrutiny rather than being recognised as a simple approach with real practical potential. It is the very lack of sophistication of our preferred version of this general approach, yet potentially yielding outstanding performance, that we feel is its strength and appeal. The interested reader accessing Jones and Henderson (2005) should be aware that it assumes a degree of novelty to its work that is actually unsupportable, although we still wish to record our debt to the exceptional doctoral work of Storvik (1999). A referee has also kindly mentioned earlier relevant doctoral work of Scott (1976) and Kim (1995), each of which works with $\widehat{f}_{\mathbf{w}}$ and notes the sparsity of the weight vector $\mathbf{w}$ when it is well chosen.

2. The estimator

Set $\mathbf{w} = \mathbf{1}$, $\mathbf{b} = h\mathbf{1}$ in (1) and consider h to be a specified value. Then

$$\hat{f}_{\mathbf{m}}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - m_i).$$

Maximising likelihood, choose $\mathbf{m}^* = \text{argmax}_{\mathbf{m}} \sum_{k=1}^{n} \log \hat{f}_{\mathbf{m}}(X_k)$. Straight away specifying $K = \phi$, the standard normal density, the estimating equations are

$$\sum_{k=1}^{n} \frac{\phi_h'(X_k - m_\ell)}{\hat{f}_{\mathbf{m}}(X_k)} = \sum_{k=1}^{n} \frac{(X_k - m_\ell)\phi_h(X_k - m_\ell)}{\hat{f}_{\mathbf{m}}(X_k)} = 0, \qquad \ell = 1, ..., n,$$

which immediately yields a beautifully simple iterative scheme for computing $\mathbf{m}^*$:

$$m_\ell^{\text{new}} = \left\{ \sum_{k=1}^{n} t_{\ell,k} \right\}^{-1} \sum_{k=1}^{n} X_k t_{\ell,k} \quad \text{with} \quad t_{\ell,k} = \frac{\phi_h(X_k - m_\ell^{\text{old}})}{\hat{f}_{\mathbf{m}^{\text{old}}}(X_k)}.$$

Finally, use $\widehat{f}_{\mathbf{m}*}$. The natural starting point for this iteration, which we always use, is $\mathbf{m} = \mathbf{X}$. The weights are all nonnegative and so each calculated m_ℓ falls within the range of $\mathbf{X}$ at all iterations.

This algorithm is, of course, an EM algorithm (e.g. Laird, 1978, DerSimonian, 1986, McLachlan and Peel, 2000). It is therefore guaranteed to increase the likelihood at each iteration and thus to converge to a local maximum of the likelihood. In our implementation of this algorithm, we deemed convergence to have occurred when the average absolute difference between the parameter values at the current and previous iterations was less than 10^{-C} for some appropriate value of C. Most of our simulations concern $n = 100$ in which case we took $C = 5$. Typical numbers of iterations to convergence were then 100–500 for $\widehat{f}_{\mathbf{m}}$. The effect on our performance measure of changing C was not large. In further simulations of the performance of $\widehat{f}_{\mathbf{m}}$ when $n = 500$, we reduced C from 5 to 3. The number of iterations was then in the tens rather than the hundreds.

The NPMLE of $g(\mu)$ in the location mixture $\int \phi_h(x - \mu)g(\mu)d\mu$ turns out to be a discrete distribution with support of size $k \leq n$ (Laird, 1978, Lindsay, 1993, 1995). The NPMLE Gaussian mixture distribution must then coincide with $\widehat{f}_{\mathbf{m},\mathbf{w}}$ (provided the algorithm used successfully maximises the likelihood). See also Proposition 1 of Geman and Hwang (1982). The estimator $\widehat{f}_{\mathbf{m}}$ can be thought of as an approximation to $\widehat{f}_{\mathbf{m},\mathbf{w}}$ which yields slightly

different results: allowing locations m to be coincident and/or weights w to be zero as necessary, notice that the actual NPMLE allows estimation of both $w_1, ..., w_n$ and $m_1, ..., m_n$ while $\widehat{f}_{\mathbf{m}}$ fixes $w_1 = ... = w_n = 1$ and varies just $m_1, ..., m_n$. For the extent of the difference in practice, see Jones and Henderson (2005).

Moreover, normal mixture NPMLE solutions have been observed to be sparse in their number of discrete support points (Laird, 1978, p. 809, Geman and Hwang, 1982, p. 411) and these sparseness properties transfer to $\widehat{f}_{\mathbf{m}}$. The beauty of this is that the form of $\widehat{f}_{\mathbf{m}}$ can then be used to suggest a much simpler parametric finite mixture model for the data. In this way, there is much added value in modelling terms in employing $\widehat{f}_{\mathbf{m}}$ over $\widehat{f}$.

It is also gratifying to find, empirically, that $\widehat{f}_{\mathbf{m}}$ seems not to change abruptly anywhere as h increases but rather changes smoothly with h in much the way one is familiar with from experience with $\widehat{f}$.

We have also investigated the use of least squares cross-validation (LSCV) to select h for $\widehat{f}_{\mathbf{m}}$. That is, h was chosen to minimise

$$\int \{\widehat{f}_{\mathbf{m}}(x;h)\}^2 dx - 2n^{-1}\sum_{i=1}^{n} \widehat{f}_{\mathbf{m},-i}(X_i;h) \tag{2}$$

where $\widehat{f}_{\mathbf{m}}(\cdot;h)$ is the density estimate (with bandwidth h) derived from the full data set, and $\widehat{f}_{\mathbf{m},-i}(\cdot;h)$ is the density estimate (with bandwidth h) based on all datapoints except X_i.

3. Simulation study

3.1. Simulation set-up

We pattern our simulation study after that of Jones and Signorini (1997) who in turn utilise the first 10 densities suggested as a simulation testbed by Marron and Wand (1992). The densities, all built from normal mixtures, are referred to as 1: Gaussian, 2: Skewed unimodal, 3: Strongly skewed, 4: Kurtotic unimodal, 5: Outlier, 6: Bimodal, 7: Separated bimodal, 8: Skewed bimodal, 9: Trimodal, 10: Claw.

To deal with the bandwidth selection problem (selection of an appropriate value for the overall value h), we first provide a "best possible" analysis in the sense that we empirically approximately compute the value of h that minimised the integrated squared error (ISE) between the estimated density

and the true density. In this way, we decouple the potential capability of each density estimator from the thorny issue of empirical bandwidth selection. That thorny issue is taken up, somewhat unsatisfactorily it turns out, via LSCV in Section 3.4.

Density estimates were calculated on an equally spaced grid of 301 points on $[-3, 3]$. The following rule was used to approximate the ISE:

$$\mathrm{ISE}(\widehat{f}) \simeq \frac{1}{50} \sum_{j=1}^{301} \left\{ \widehat{f}(y_j) - f(y_j) \right\}^2,$$

where $y_j = -3 + (j-1)/50$. Minimisation over h was performed over a carefully chosen grid of values. Our procedure was: (i) to start with an equally-spaced grid of 12 points from 0.1 to 1.2. (If the optimal h was not in this range, a different grid was tried); (ii) based on the results of the first grid, a second, finer, grid needed to be used for some densities. These finer grids were nearly always between 9 and 13 point grids but not always equally-spaced; (iii) this ad hoc process was repeated — rarely beyond three different grids — until a 'satisfactory' grid of values was obtained and used thereafter.

Our main simulation study concerns samples of size $n = 100$, replicating the experiment 1000 times. This is the subject of Section 3.2. A smaller simulation study when $n = 500$ (also with 1000 replications) is reported in Section 3.3. The case $n = 100$ is revisited in Section 3.4 where the ISE-optimal bandwidth selector is replaced by the LSCV bandwidth selector. The normal kernel is used in all cases.

3.2. Simulation results, $n = 100$

The results of our simulations when $n = 100$, which will be found in Table 1, follow Jones and Signorini (1997) in giving the mean and standard error of the minimised ISE ($\times 10^5$) calculated over the 1000 simulated datasets for each estimator/density combination. Also given in Table 1 is the median percentage reduction of the minimised ISE for the estimator $\widehat{f}_{\mathbf{m}}$ compared with that of $\widehat{f}$. Clearly, $\widehat{f}_{\mathbf{m}}$ proves to be superior to $\widehat{f}$ in 8 of the 10 cases, often considerably so (although once — for model 8 — the improvement is not statistically significant). Its only significant disimprovement is for model 3, the strongly skewed density. Our results are more favourable towards $\widehat{f}_{\mathbf{m}}$ compared with $\widehat{f}$ than were simulations alluded to by Geman and Hwang

Table 1
Means (standard errors in parentheses) of minimized $ISE \times 10^5$ for samples of size $n = 100$ from each of the first ten Marron-Wand densities over 1000 simulations. The median % reduction is given on the second line for $\widehat{f}_{\mathbf{m}}$. ISE-optimal bandwidths are used.

Estimator	Density				
	Gaussian	Skewed unimodal	Strongly skewed	Kurtotic unimodal	Outlier
$\widehat{f}$	493 (13)	779 (18)	4165 (49)	4044 (56)	5188 (121)
$\widehat{f}_{\mathbf{m}}$	165 (6)	442 (12)	4645 (53)	3526 (48)	2212 (67)
	72.5	45.1	−11.8	11.6	59.5
	Bimodal	Separated bimodal	Skewed bimodal	Trimodal	Claw
$\widehat{f}$	706 (13)	1099 (19)	903 (14)	860 (14)	3540 (34)
$\widehat{f}_{\mathbf{m}}$	537 (13)	633 (15)	938 (16)	791 (14)	3429 (37)
	28.5	45.5	−1.6	9.1	3.1

(1982, p. 410); the reason, we suspect, is that Geman and Hwang used the same bandwidth for both estimators.

A parallel set of results for a variety of different (fourth order bias) estimators is given in Jones and Signorini (1997, Table 1; henceforth JST1). We can compare the results of Table 1 with those of JST1, bearing in mind that the two refer to different sets of simulations, that JST1 uses the biweight kernel while Table 1 uses the normal, and that in the current experiment, we may not have computed the ISE-optimal bandwidth to such a high degree of accuracy. Comparison of results for $\widehat{f}$ in Table 1 and in JST1 shows the two to be comparable to the levels of accuracy required to make general

claims. In particular, we can compare the performance of $\widehat{f}_{\mathbf{m}}$ in Table 1 with that of its "asymptotics-based" counterpart, Jones and Signorini's (1997) implementation of the variable location estimator ($\widehat{f}_6$ in JST1): $\widehat{f}_{\mathbf{m}}$ shows considerably better performance than JST1's $\widehat{f}_6$ except for slightly inferior performance for model 3 and, perhaps, comparable performance for model 8. We also note in Jones and Henderson (2005) that the fully iterated estimator $\widehat{f}_{\mathbf{m}}$ generally outperforms its alternative version utilising just one iteration of the algorithm.

Finally, how does $\widehat{f}_{\mathbf{m}}$ compare with the best of the single-bandwidth higher order bias estimators in JST1? (There are also two-bandwidth versions of some estimators in JST1, but Jones and Signorini, 1997, Section 6, are suspicious of their worth because of the extra difficulties in realising any potential improvements in practice involving data-driven bandwidth selection.) This best estimator appears to be what JST1 calls $\widehat{f}_{3,1}$, the multiplicatively bias corrected density estimator of Jones, Linton and Nielsen (1995). And, consistently across all densities except numbers 3 (barely significantly) and 8 (insignificantly), $\widehat{f}_{\mathbf{m}}$ outperforms $\widehat{f}_{3,1}$, albeit not by huge amounts. This is consistent with Jones and Signorini's observation that the variable location estimator in JST1, $\widehat{f}_6$ is "a little behind $\widehat{f}_{3,1}$" and the observation made about the comparative performance of $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_6$ above.

3.3. Simulation results, $n = 500$

In order to check that our claims of good performance of $\hat{f}_{\mathbf{m}}$ are not confined, for some reason, to $n = 100$, we repeated the above experiment for $n = 500$. The results are in Table 2. Whether measured by mean or median reduction, the performance of $\widehat{f}_{\mathbf{m}}$ relative to that of $\widehat{f}$ is enhanced over that for $n = 100$ in every single case. (This even translates a slightly worse relative performance for model 8, the skewed bimodal, when $n = 100$ into a slightly better relative performance when $n = 500$.) Also, when $n = 500$, $\widehat{f}_{\mathbf{m}}$ continues to perform generally better than the Jones, Linton and Nielsen estimator (in JST1), perhaps even a little enhanced relatively in most cases, although slightly less well for a few. By the way, we have evidence that the relaxing of the convergence criterion for $n = 500$ mentioned in Section 2 has made the ISE values for $\widehat{f}_{\mathbf{m}}$ a little larger than they might be, but this effect is generally small.

Table 2
Means (standard errors in parentheses) of minimized $ISE \times 10^5$ for samples of size $n = 500$ from each of the first ten Marron-Wand densities over 1000 simulations. The median % reduction is given on the second line for $\widehat{f}_{\mathbf{m}}$. ISE-optimal bandwidths are used.

Estimator	Density				
	Gaussian	Skewed unimodal	Strongly skewed	Kurtotic unimodal	Outlier
$\widehat{f}$	158 (3)	247 (5)	1378 (15)	1249 (16)	1634 (35)
$\widehat{f}_{\mathbf{m}}$	35 (1)	109 (3)	1439 (14)	988 (12)	571 (16)
	84.6	59.7	−2.7	21.0	67.9
	Bimodal	Separated bimodal	Skewed bimodal	Trimodal	Claw
$\widehat{f}$	230 (4)	325 (5)	301 (5)	284 (4)	1117 (11)
$\widehat{f}_{\mathbf{m}}$	128 (3)	132 (3)	275 (5)	258 (4)	884 (10)
	48.5	63.1	10.3	10.3	21.6

3.4. Simulation results, $n = 100$, LSCV bandwidth selection

After computing formula (2) on an equi-spaced 12-point grid of values of $\log(h)$ from -5 to 0.5, a new initial grid was formed, tailored to each specific model. The LSCV bandwidth was taken to be the largest local minimiser of the LSCV function (Marron, 1993, Wand and Jones, 1995, Section 3.3), where the `optimize` function in `R` (R Development Core Team, 2008) was applied to the interval $(h_{\text{low}}, h_{\text{high}})$, these being the grid values immediately below and above the (tailored) grid value corresponding to the largest minimiser. The same procedure was applied to $\widehat{f}$ as to $\widehat{f}_{\mathbf{m}}$. Results are in Table 3.

Table 3
Means (standard errors in parentheses) of minimized $ISE \times 10^5$ for samples of size $n = 100$ from each of the first ten Marron-Wand densities over 1000 simulations. The median % reduction is given on the second line for $\widehat{f}_{\mathbf{m}}$. LSCV bandwidths are used.

Estimator	Density				
	Gaussian	Skewed unimodal	Strongly skewed	Kurtotic unimodal	Outlier
$\widehat{f}$	834 (25)	1246 (35)	5136 (73)	5013 (82)	8380 (341)
$\widehat{f}_{\mathbf{m}}$	353 (14) 63.9	834 (23) 31.1	6633 (95) −28.5	5154 (99) 2.0	7909 (361) 30.9
	Bimodal	Separated bimodal	Skewed bimodal	Trimodal	Claw
$\widehat{f}$	1025 (25)	1456 (33)	1217 (21)	1176 (22)	4971 (38)
$\widehat{f}_{\mathbf{m}}$	1856 (6) −112.2	921 (24) 40.5	2108 (12) −84.0	1179 (21) −6.4	5533 (12) −7.4

First, performance is, naturally, rather worse for both estimators than it was for ISE-optimal bandwidths. Second, performance of $\widehat{f}_{\mathbf{m}}$ is, in terms relative to that of $\widehat{f}$, consistently a bit worse for LSCV bandwidth selection than it was for ISE-optimal bandwidth selection. (In one case, model 6, bimodal, there is particularly strong disimprovement for $\widehat{f}_{\mathbf{m}}$ relative to $\widehat{f}$.) The result is just three clear 'wins' for $\widehat{f}_{\mathbf{m}}$, four for $\widehat{f}$ and three cases where the difference is not statistically significant.

Further investigation by scatterplot and other pictorial devices (not shown) of the joint behaviour of the LSCV and ISE-optimal bandwidths in the simulations was also done. This showed, in most cases, a reasonably good centring

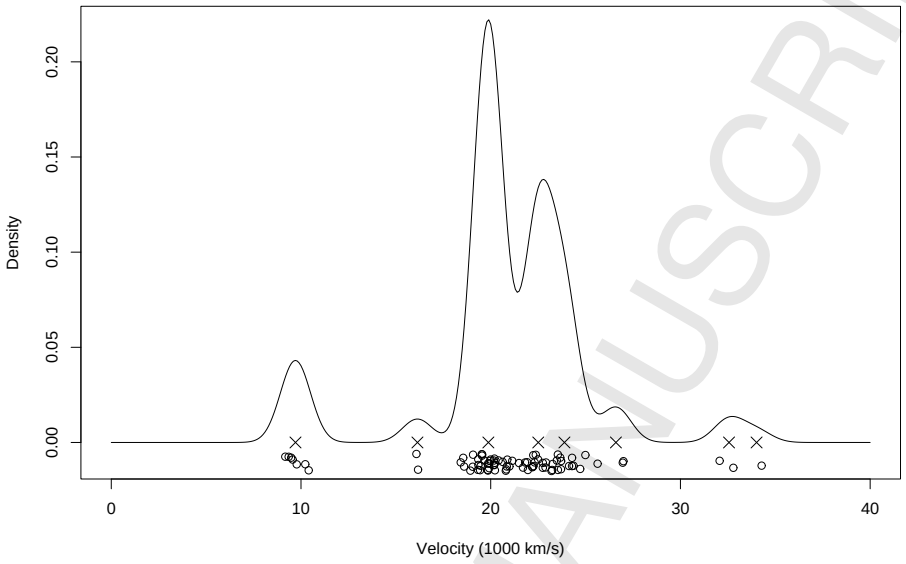
of the distribution of the former on the latter but a particularly large degree of variability in the LSCV bandwidth; the two bandwidths were not strongly correlated. This behaviour is essentially as is familiar from the ordinary kernel density estimation case. For the more structured densities (especially numbers 6, 8 and 10), LSCV bandwidths tended to be much larger than the ISE-optimal bandwidths.

The hunt for better bandwidth selectors for use with $\widehat{f}_{\mathbf{m}}$ should, therefore, be pursued. The potential improvement to be afforded by $\widehat{f}_{\mathbf{m}}$ is not realised with LSCV bandwidth selection. In saying this, we are at odds with Geman and Hwang's (1982, p.412) statement, made after "experiment[ing] extensively" in the sieve context, that "cross-validation is often a strikingly effective means of choosing an appropriate degree of smoothing". This may partly be a matter of semantic degree in the use of the word "effective". Our own efforts to find an empirical relationship between bandwidths appropriate to $\widehat{f}_{\mathbf{m}}$ and to $\widehat{f}$ and thence to provide a 'rule-of-thumb' approach to bandwidth selection, although initially promising, eventually proved fruitless, so no further details are given.

4. Example

As an illustrative example of the methodology, we estimate the density of velocities associated with $n = 82$ galaxies in the Corona Borealis region of the sky. These data, due to Postman, Huchra and Geller (1986), are as given and studied by Roeder (1990). (We join Roeder in working in units of 1000km/s.) We use $\widehat{f}_{\mathbf{m}}$ with least squares cross-validation selection of h, Section 3.4 notwithstanding. The cross-validated value of $h = 0.79$. The data and resulting density estimate are shown in Fig. 1. The density estimate in Fig. 1 displays six modes. For convenience, call them modes 1 to 6 reading from left to right. In comparison with the least squares cross-validated kernel density estimate given in Roeder (1990, Fig. 1), our estimate is a little less smooth: in Roeder's estimate, modes 3 and 4 are not separated, there is but the vaguest suggestion of mode 2 and none of mode 5, and there is an extra small mode to the left of mode 1. The kernel density estimate with (smaller) bandwidth selected by the Sheather and Jones (1991) method (not shown) is more similar to Fig. 1; the differences are only less accentuation of modes 3 and 4 and the division of mode 6 into two modes. Fig. 1 is pretty similar to Fig. 5(c) of Roeder (1990) where a parametric six component normal mixture (with equal variances) has been fitted. It differs relatively little from Roeder's

Fig. 1. The variable location density estimate $\widehat{f}_{\mathbf{m}}$ with $h = 0.79$ for the galaxy velocity data. The datapoints are plotted as vertically jittered circles. The eight distinct locations involved in $\widehat{f}_{\mathbf{m}}(\cdot; 0.79)$ are shown by crosses.



(1990, Fig. 7) preferred normal mixture-based density estimate which has five modes, mode 2 not being present. However, Roeder's five mode density corresponds to a mixture model with some 17 components, although she does state that "this is partly an artifact of [her particular implementation of her] algorithm".

However, $\widehat{f}_{\mathbf{m}}$ has, without any prior specification of, or any complex procedure directly selecting, the number of components, reduced to a relatively simple normal mixture model anyway. This is an illustration of the general point made in Section 2. The normal mixture has eight components; their locations are plotted on Fig. 1 and their weights and precise locations are given in Table 4. Modes 4 and 6 are each made up of two normal components suggesting skewness. Of course, arguments about mode 6 are meaningless, there being only three datapoints involved, but that of mode 4 is based on 31 of the 82 datapoints and hence may be more interesting.

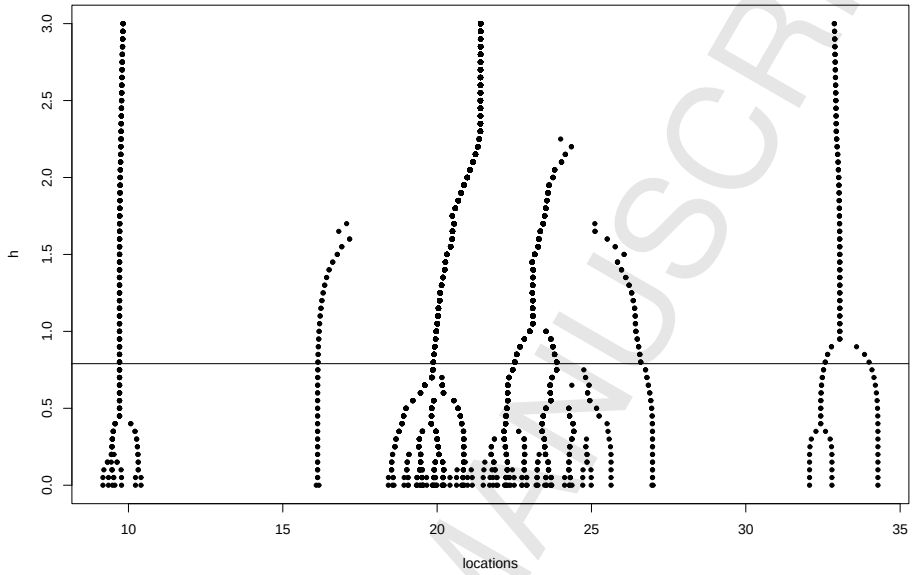
Table 4
The normal mixture components associated with $\widehat{f}_{\mathbf{m}}(\cdot; 0.79)$ for the galaxy velocity data, $n = 82$. Each has variance $0.79^2 = 0.624$.

Component	Weight ($\times 82$)	Mean
1	7	9.710
2	2	16.138
3	36	19.876
4	19	22.507
5	12	23.885
6	3	26.599
7	2	32.561
8	1	34.014

More information on the points of support of $\widehat{f}_{\mathbf{m}}$ is provided by plotting the maximum likelihood locations as a function of increasing bandwidth. Such a plot is strongly reminiscent of the 'mode tree' of Minnotte and Scott (1993) in which the modes of the density estimate, rather than the underlying mixture locations, are plotted as a function of h. This aesthetically pleasing plot, shown in Fig. 2, contributes evidence to our claim (Section 2) that $\widehat{f}_{\mathbf{m}}$ changes smoothly with h; further investigation of, for example, the region around $h = 1.7$ where two locations disappear at essentially the same time shows that even this still corresponds to only a very moderate change in the density estimate with h. A referee has suggested that such a plot might have more sophisticated uses (e.g. to do with bandwidth selection) but these have not yet been explored.

Our illustration of the use of $\widehat{f}_{\mathbf{m}}$ stops there but further modelling of the data might start with the eight component mixture model we have found and work parametrically towards a simpler model (e.g. without two components for mode 6).

Fig. 2. Plot of maximum likelihood locations obtained for different values of h from 0 to 3 when applying $\widehat{f}_{\mathbf{m}}$ to the galaxy velocity data. The data are themselves plotted, masquerading as the locations corresponding to $h = 0$. The horizontal line is drawn at the LSCV selection of $h = 0.79$.



5. Conclusions

There has been much interest over the years in sophistications of the basic kernel density estimator. A major strand of work has focussed on single adaptations, such as higher order kernels, variable bandwidths, locations or weights, transformations and multiplicative corrections designed to improve performance. Jones and Signorini (1997) reviewed and compared many of these and came to the conclusion that "It remains debatable, however, as to whether even the best methods give worthwhile improvements, at least for small-to-moderate sample exploratory purposes". While they argued at the time that " 'failings' of such estimators thus are often due not to substandard methodology, but rather to limitations in the information available in the data", it has since transpired that certain methodological improvements remain possible. Two particular negatives associated with the types

of method considered by Jones and Signorini are an overreliance first on asymptotics and second on reducing bias with no particular regard paid to variance. Maximum likelihood kernel/convolution sieve density estimation seeks to sweep these considerations aside in one fell swoop.

Another strand of work has sought to try to reduce the complexity and improve the interpretability of kernel density estimates by (usually complicated) schemes involving removing/reweighting kernels while not changing the density estimate appreciably. Examples include Marchette *et al.* (1996), Priebe and Marchette (2000) and Scott and Szewczyk (2001). We feel that this work, too, is somewhat undermined by the simplicity of the action of kernel convolution sieve density estimation in automatically producing interpretable solutions of the type desired. That said, a referee has suggested that the current work could serve to provide an input mixture to, say, the filtered kernel density estimator and that this added complexity may, in some cases, be warranted.

We repeat our admission of a failure to obtain the kinds of theoretical results we would like in terms of asymptotic bias and variance. We hope that publication of this paper will inspire such theory to be developed. Not only would it be informative about the quality of maximum likelihood kernel density estimation per se, but it should give assistance concerning the bandwidth selection problem, thereby replacing the computationally clumsy and, as we have shown, rather ineffective, cross-validation procedure. As suggested by a referee, successful efforts at speeding up the computation of $\widehat{f}_{\mathbf{m}}$ would also be a valuable contribution.

To finish, though, we re-emphasise that the simple location-only maximum likelihood kernel density estimator or kernel convolution sieve, $\widehat{f}_{\mathbf{m}}$, seems to provide an especially promising framework for further practical utilisation and development.

Acknowledgement

This work started on a visit of the first author to the Chinese University of Hong Kong in 2001. We are most grateful to Jianqing Fan for his hospitality and for his continuing interest, encouragement and efforts (along with those of one of his colleagues). We are also extremely grateful to anonymous referees of Jones and Henderson (2005) who pointed out our errors in terms of precedents in the literature and to the referees of the current version of the paper for prompting further improvements to it.

References

Abramson, I.S., 1982. On bandwidth variation in kernel estimates — a square root law. Ann. Statist. 9, 168–176.

Choi, E., Hall, P, 1999. Data sharpening as a prelude to density estimation. Biometrika 86, 941–947.

DerSimonian, R., 1986. Algorithm AS 221. Maximum likelihood estimation of a mixing distribution. Appl. Statist. 35, 302–309.

Geman, S., Hwang, C.R., 1982. Nonparametric maximum likelihood estimation by the method of sieves. Ann. Statist. 10, 401–414.

Ghosal, S., van der Vaart, A.W., 2001. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann. Statist. 29, 1233–1263.

Hall, P., Turlach, B.A., 1999. Reducing bias in curve estimation by use of weights. Comput. Statist. Data Anal. 30, 67–86.

Jones, M.C., Henderson, D.A., Maximum likelihood kernel density estimation. (Technical report 01/05, Department of Statistics, The Open University, UK, 2005. Available from `http://statistics.open.ac.uk/TechnicalReports/TechnicalReportsIntro.htm`)

Jones, M.C., Linton, O., Nielsen, J.P., 1995. A simple and effective bias reduction method for kernel density estimation. Biometrika 82, 327–338.

Jones, M.C., Signorini, D.F., 1997. A comparison of higher order bias kernel density estimators. J. Amer. Statist. Assoc. 92, 1063–1073.

Kim, D., 1995. Least squares mixture decomposition estimation. (Ph. D. thesis, Virginia Polytechnic Institute and State University)

Laird, N.M., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. J. Amer. Statist. Assoc. 73, 805–811.

Lindsay, B.G., 1983. The geometry of mixture likelihoods: a general theory. Ann. Statist. 11, 86–94.

Lindsay, B.G., 1995. Mixture Models: Theory, Geometry and Applications. Institute of Mathematical Statistics, Hayward, CA.

Marchette, D.J., Priebe, C.E., Rogers, G.W., Solka, J.L., 1996. Filtered kernel density estimation. Comput. Statist. 11, 95–112.

Marron, J.S., 1993. Contribution to the discussion of "Practical performance of several data driven bandwidth selectors" by Park and Turlach. Comput. Statist. 8, 17–19.

Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Statist. 20, 712–736.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Minnotte, M.C., Scott, D.W., 1993. The mode tree: a tool for visualization of noparametric density features. J. Comput. Graphical Statist. 64, 141–157.

Postman, M., Huchra, J.P., Geller, M.J., 1986. Probes of large-scale structures in the Corona Borealis region. Astronom. J. 92, 1238–1247.

Priebe, C.E., Marchette, D.J., 2000. Alternating kernel and mixture density estimates. Comput. Statist. Data Anal. 35, 43–65.

R Development Core Team, 2008. The R project for statistical computing. (`http://www.r-project.org`)

Roeder, K., 1990. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. J. Amer. Statist. Assoc. 85, 617–624.

Samiuddin, M., el-Sayyad, G.M., 1990. On nonparametric kernel density estimates. Biometrika 77, 865–874.

Scott, D.W., 1976. Nonparametric probability density estimation by optimization theoretic techniques. (Ph. D. thesis, Rice University)

Scott, D.W., Szewczyk, W.F., 2001. From kernels to mixtures. Technometrics 43, 323–335.

Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. Ser. B 53, 683–690.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Storvik, B.E., 1999. Contributions to nonparametric curve estimation. (Dr. Scient. thesis, University of Oslo)

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman and Hall,

London.